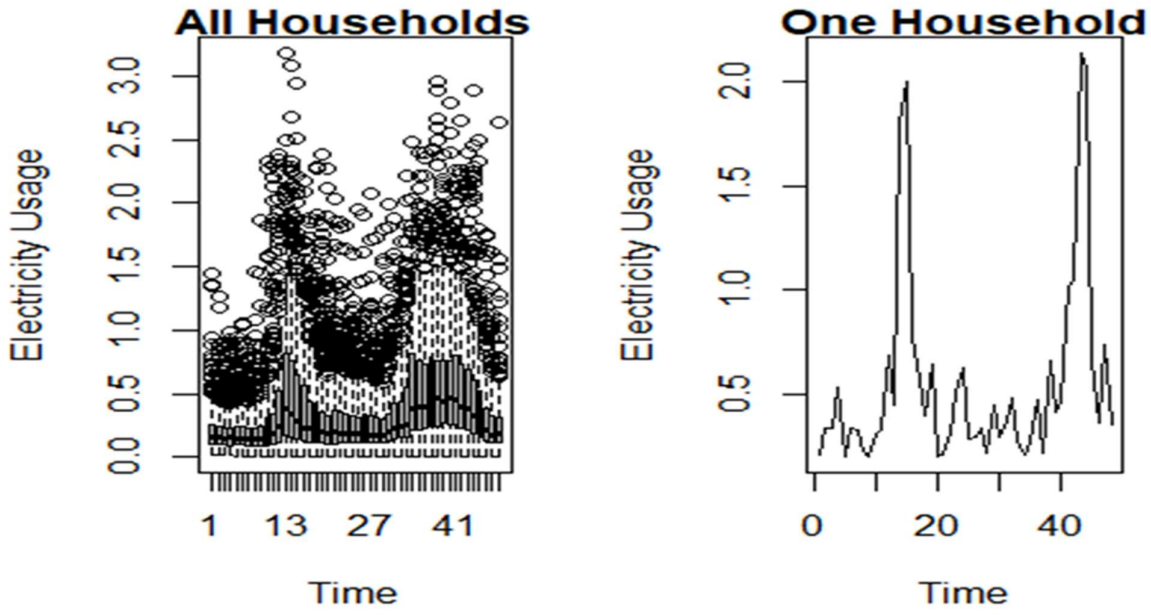
ELECTRICITY DATA MODELLING

1, Examine the boxplot and briefly discuss the aggregated pattern of electricity usage.



From the boxplot shown above, I found the following information:

1, There are two peak time ranges in electricity usage, one in the morning around 6 am to 8 am, and another one in the evening around 7 pm to 9 pm. These reflect people's normal life rules.

2, The electricity usage gradually increases from the midnight to the morning (breakfast time), and it gradually decreases from morning to the lunchtime (looks like most people have no time to go back home during the lunchtime), and then it gradually increased from the lunchtime to the evening (dinner time), after the dinner time, the electricity usage gradually decreased until another midnight time.

3, Most people use electricity in accordance with the number 2 rules listed above, but there are some exceptions, we can think of these small exceptions as outliers.

2, (a) Construct a table of explanatory variables, where each row is a household, and each column is an explanatory variable that you have constructed from the original usage data.

My R code for constructing the table of explanatory variables is as follows:

```
setwd("C:/Users/zizhe/Desktop)

p <- read.csv('power.csv',header=FALSE)

 rownames(p) <- p[,1]

p <- p[,2:ncol(p)]

colnames(p) <- seq(from=1,to=48,by=1)
```

```
meancol<-apply(p,1,mean)
 varcol<-apply(p,1,var)
maxcol<-apply(p,1,max)
 mincol<-apply(p,1,min)

usage0006<-apply(p[,1:12],1,sum)
usage0618<-apply(p[,13:36],1,sum)
usage1824<-apply(p[,37:48],1,sum)

newtable<-cbind(meancol,varcol,maxcol,mincol,usage0006,usage0618,usage1824)
colnames(newtable)<-c("mean","var","max","min","usage 0am-6am","usage 6am-6pm","usage 6pm-0am")
newtable<-as.data.frame(newtable)
```

The newtable screenshot is as below:

```
> newtable<-as.data.frame(newtable)
> newtable
                      mean          var   max   min usage 0am-6am usage 6am-6pm usage 6pm-0am
Reading.207860297  0.39243750 0.0761882088 1.164 0.153         3.631         8.546         6.660
Reading.207860316  0.55677083 0.2358126059 2.140 0.199         4.045        12.491        10.189
Reading.207860322  0.46779167 0.2594152748 2.251 0.080         2.502         6.747        13.205
Reading.207860324  0.25237500 0.0509443670 1.094 0.041         1.313         6.753         4.048
Reading.207860359  0.39722917 0.1144092868 1.568 0.084         1.685        11.706         5.676
Reading.207860373  0.87422917 0.4158974570 2.653 0.292         7.627        18.712        15.624
Reading.207860393  0.34962500 0.0669094309 1.116 0.074         2.685         8.986         5.111
Reading.207860395  0.59133333 0.2398432482 2.095 0.182         3.509        15.027         9.848
Reading.207860402  0.27260417 0.0393101591 0.912 0.079         2.531         4.677         5.877
Reading.207860439  0.60116667 0.1524692482 1.825 0.210         4.334        14.983         9.539
Reading.207860447  0.40479167 0.1215220833 1.758 0.073         3.080         9.257         7.093
```

(b) Describe the functions that you have constructed to create these explanatory variables, and in particular discuss how you have described (represented) the temporal structure of usage.

The 4 functions I used for this task are mean, variance, max, min.

- the mean function is very important, the purpose of the mean value is to represent the average level of electricity usage in each household each day. It is a statistic that describes the degree of aggregation of data. The mean value can reflect the daily electricity usage of the household, and different marketing strategies can be formulated according to the daily electricity usage of the different households.

- the variance function is also important, The variance is to see how discrete the sample is. Variance is used to calculate the difference between each variable and the population mean. When the variance value is large, it indicates that the data distribution of household's electricity usage is

scattered and the data fluctuation is large; when the variance value is small, it indicates that the data distribution of household's electricity usage is concentrated, and the data fluctuation is small.

- The max function and the min function are important, The maximum value and minimum value reflect the maximum and minimum electricity usage of each household at a certain time within 24 hours every day. They are the true reflection of the data range of the household's daily electricity usage.

Regarding the temporal structure of usage, I divided the time spent using electricity into three parts, The 3-time ranges are 0am-6am, 6am-6pm, and 6pm-0am. the total electricity usage in each of these 3-time frames is important for data clustering because we can group households that have similar electricity usage in each of the 3-time ranges together, and offer them a special package based on their electricity usage amount. For example, the marketing plan can offer a cheaper package price to the households which are using more electricity in the evening, this marketing plan can attract more new potential customers.

```
> newtable<-as.data.frame(newtable)
> newtable
                      mean          var   max   min usage 0am-6am usage 6am-6pm usage 6pm-0am
Reading.207860297  0.39243750 0.0761882088 1.164 0.153         3.631         8.546         6.660
Reading.207860316  0.55677083 0.2358126059 2.140 0.199         4.045        12.491        10.189
Reading.207860322  0.46779167 0.2594152748 2.251 0.080         2.502         6.747        13.205
Reading.207860324  0.25237500 0.0509443670 1.094 0.041         1.313         6.753         4.048
Reading.207860359  0.39722917 0.1144092868 1.568 0.084         1.685        11.706         5.676
Reading.207860373  0.87422917 0.4158974570 2.653 0.292         7.627        18.712        15.624
Reading.207860393  0.34962500 0.0669094309 1.116 0.074         2.685         8.986         5.111
Reading.207860395  0.59133333 0.2398432482 2.095 0.182         3.509        15.027         9.848
Reading.207860402  0.27260417 0.0393101591 0.912 0.079         2.531         4.677         5.877
Reading.207860439  0.60116667 0.1524692482 1.825 0.210         4.334        14.983         9.539
Reading.207860447  0.40479167 0.1215220833 1.758 0.073         3.080         9.257         7.093
```

I would like to use one household as an example to explain. From the screenshot above, we can see the first row is a record of electricity usage for one household. Between the time 0am-6am, the electricity usage value totally is 3.631, between the time 6am-6pm, the electricity usage value totally is 8.546, and between the time 6pm-0am, the electricity usage value totally is 6.660. We can get the conclusion that during the late-night time between 0am-6am, the electricity usage is low. During the evening time until the mid-night time, 6pm-0am, the electricity usage somehow high, and during the day time 6am-6pm, the electricity usage is the highest. But not all the households can follow this electricity usage pattern, some households maybe use more electricity at 6pm-0am compared with the time range 6am-6pm.

3, Apply K-means clustering with 6 centers to your final explanatory data table. Produce two figures: one with boxplots showing the final clustering patterns of usage for each of the six clusters, the second as 6-line plots showing the mean usage pattern for each timestep for each cluster. See Figures 1 and 2 below for an example Comment on how the patterns of usage vary.

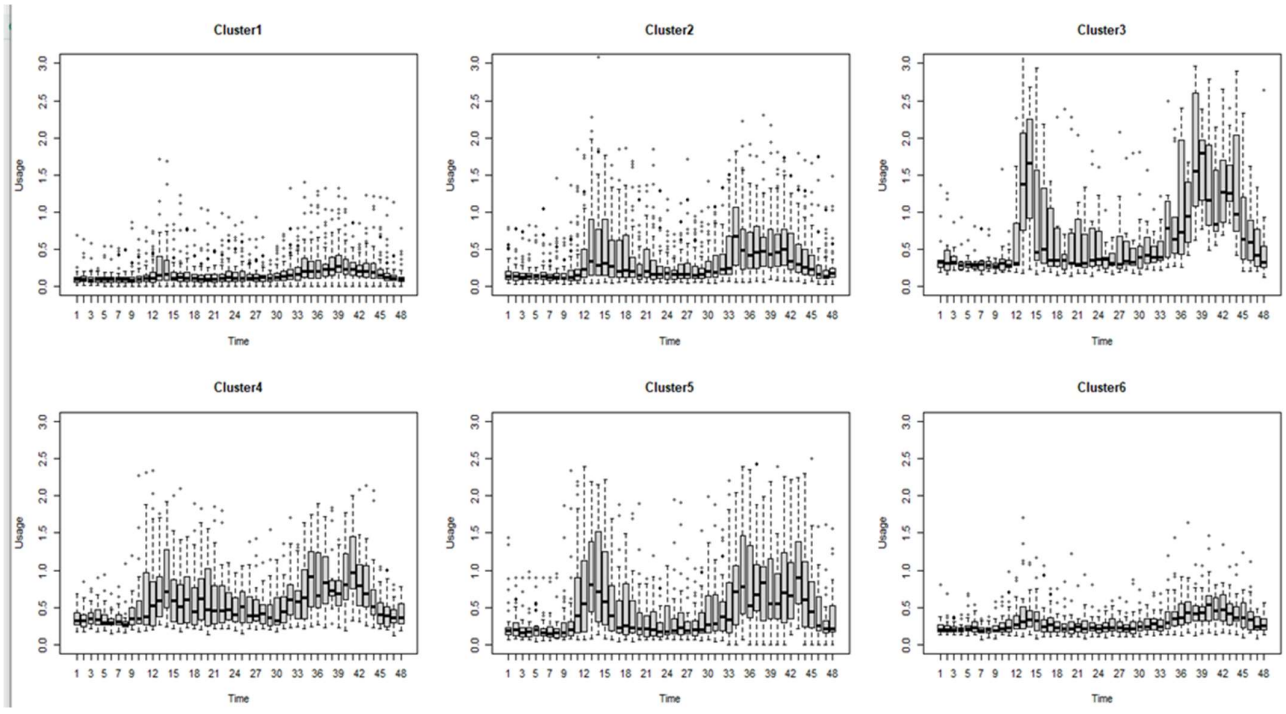
The R code for producing the clusters is as follows:

```
cl <- kmeans(scale(newtable), centers = 6)

cl1<-p[cl$cluster==1,]

cl2<-p[cl$cluster==2,]
```

```
cl3<-p[cl$cluster==3,]
cl4<-p[cl$cluster==4,]
cl5<-p[cl$cluster==5,]
cl6<-p[cl$cluster==6,]
```

The R code for producing the boxplot of usage for each of the 6 clusters is as follows:

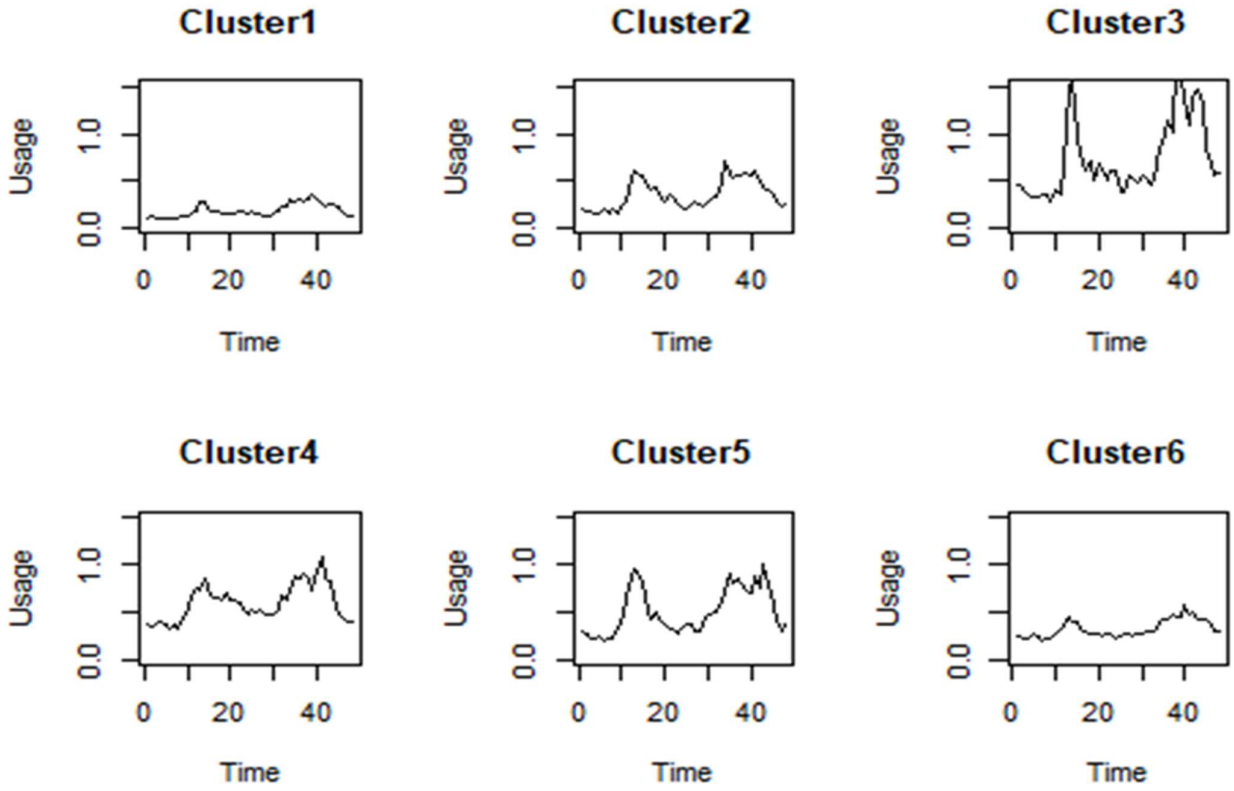
```
par(mfrow=c(2, 3))
boxplot(cl1,xlab="Time",ylab="Usage", main="Cluster1",ylim=c(0.0,3.0))
boxplot(cl2,xlab="Time",ylab="Usage", main="Cluster2",ylim=c(0.0,3.0))
boxplot(cl3,xlab="Time",ylab="Usage", main="Cluster3",ylim=c(0.0,3.0))
boxplot(cl4,xlab="Time",ylab="Usage", main="Cluster4",ylim=c(0.0,3.0))
boxplot(cl5,xlab="Time",ylab="Usage", main="Cluster5",ylim=c(0.0,3.0))
boxplot(cl6,xlab="Time",ylab="Usage", main="Cluster6",ylim=c(0.0,3.0))
```



The R code for producing the mean usage for each of the 6 clusters is as follows:

```
cl1mean <- apply(cl1,2,mean)
cl2mean <- apply(cl2,2,mean)
cl3mean <- apply(cl3,2,mean)
```

```
cl4mean <- apply(cl4,2,mean)
cl5mean <- apply(cl5,2,mean)
cl6mean <- apply(cl6,2,mean)

plot(cl1mean,type="l",xlab="Time",ylab="Usage", main="Cluster1",ylim=c(0.0,1.5))
plot(cl2mean,type="l",xlab="Time",ylab="Usage", main="Cluster2",ylim=c(0.0,1.5))
plot(cl3mean,type="l",xlab="Time",ylab="Usage", main="Cluster3",ylim=c(0.0,1.5))
plot(cl4mean,type="l",xlab="Time",ylab="Usage", main="Cluster4",ylim=c(0.0,1.5))
plot(cl5mean,type="l",xlab="Time",ylab="Usage", main="Cluster5",ylim=c(0.0,1.5))
plot(cl6mean,type="l",xlab="Time",ylab="Usage", main="Cluster6",ylim=c(0.0,1.5))
```



Comment on how the patterns of usage vary.

From the screenshot above, we can find that:

1, The usage patterns in Cluster1 and Cluster6 look very similar. They both show very low usage amount, probably no people at home the whole day. But Cluster 6 looks like a little bit more usage compared with Cluster1. Households in cluster6 may have more power-consuming devices, such as multiple

refrigerators, which are more power-consuming. However, both Cluster1 and Cluster6 use significantly less power than other clusters, indicating no signs of human consumption.

2, The usage patterns in Cluster2 and Cluster4, and Cluster5 look closely. They all showed higher usage of electricity in the morning and evening. In Cluster2, probably the people amount very small, for example just a couple with a kid. The couple needs to full-time work outside and the kid needs to go to school. In Cluster4, probably there are more people living in the household. Maybe there are a lot of retired people in the house, or maybe there are a lot of kids who use video games a lot and consume a lot of electricity. In Cluster 4, the household maybe has patients in the home who use power-hungry medical devices for a long time. The electricity usage happened throughout the whole daytime. The usage of electricity in Cluster4 in the morning and evening appears to be fairly balanced. In Cluster5, the hours of high usage of electricity were longer in the evening than in the morning.

3, In Cluster3, the household electricity usage in the morning and evening is both very large, and the electricity usage time in the evening is longer than that in the morning. Probably there are a lot of people who live in this kind of household, probably someone does the home business to produce some product which needs to consume more electricity. It could also be the result of having a lot of friends over for parties every night and using a lot of power-hungry household goods, such as an electric barbecue stove. Cluster3 shows high electricity usage at all times except during sleeping hours, especially in the morning and evening.

4, Select a pair of explanatory variables (say x1 and x2) that are not highly correlated and plot x1 versus x2 for each household colouring each point by cluster number. Comment on how the clustering is related to these variables.

The R code for this task is as follows:

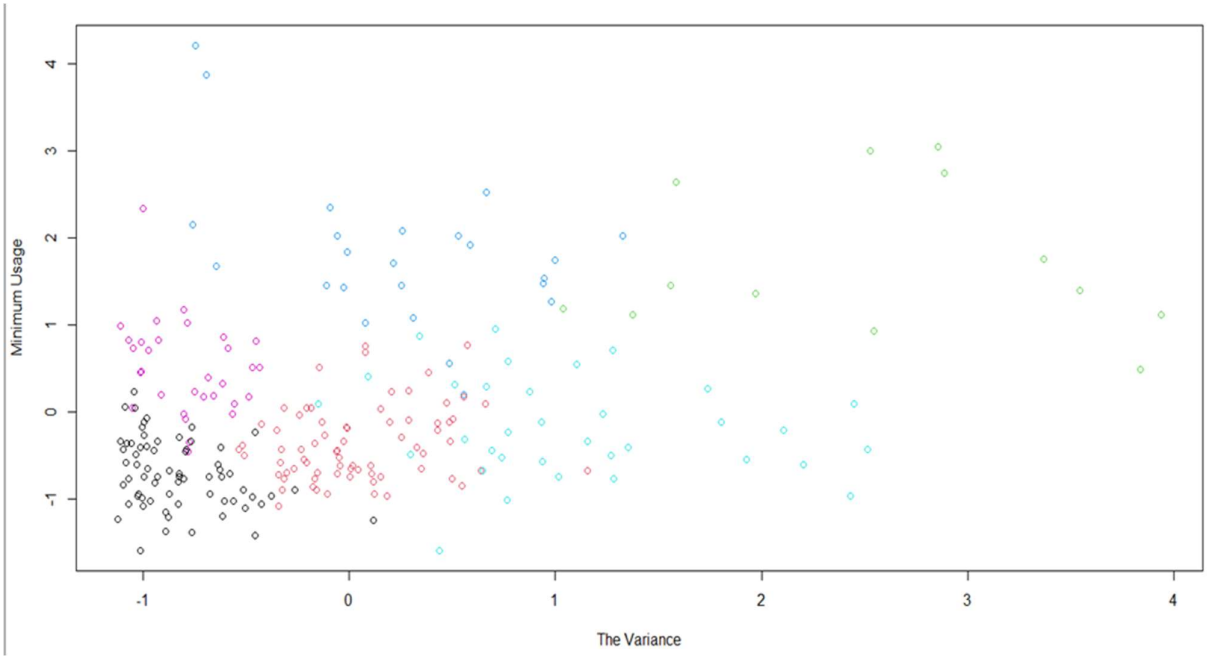
```
newtable<-as.matrix(newtable)

cor(newtable)
```

```
> cor(newtable)
                     mean       var       max       min usage 0am-6am usage 6am-6pm usage 6pm-0am
mean           1.0000000 0.7879741 0.7530107 0.6978437     0.7076315     0.9186208     0.8369370
var            0.7879741 1.0000000 0.8819162 0.3203398     0.4825414     0.6813076     0.7614171
max            0.7530107 0.8819162 1.0000000 0.3418873     0.4874088     0.6509749     0.7139231
min            0.6978437 0.3203398 0.3418873 1.0000000     0.5883457     0.5978938     0.5974534
usage 0am-6am  0.7076315 0.4825414 0.4874088 0.5883457     1.0000000     0.5370285     0.4945790
usage 6am-6pm  0.9186208 0.6813076 0.6509749 0.5978938     0.5370285     1.0000000     0.5995957
usage 6pm-0am  0.8369370 0.7614171 0.7139231 0.5974534     0.4945790     0.5995957     1.0000000
```

From the screenshot above, we can see that the min and the var are not highly correlated. Because the number 0.3203398 is the smallest one.

Plotting the min and the var:

```
par(mfrow=c(1,1))

plot(scale(newtable)[,2],scale(newtable)[,4],col = cl$cluster,xlab="The Variance",ylab="Minimum Usage")
```



Comment on how the clustering is related to these variables:

From the screenshot above, we can see that the clusters are separated out in different colors. When the values of the variance variable and the minimum usage variable are lower, the density of the cluster is higher; when the values of the variance variable and the minimum usage variable are higher, the density of the cluster is sparser and more scattered. They can largely explain those different cluster groupings.